山东大学
SHANDONG UNIVERSITY

高级机器学习

报告题目：用于心电图分类的具有内部表示连接的双峰掩蔽自编码器

姓　　名	项贤通
学　　号	202435378
学　　院	软件学院
年　　级	2024 级
专　　业	人工智能
授课老师	连莉

2024 年 11 月 16 日

摘　　要

时间序列的自监督方法被广泛应用，心电图（ECG）分类任务也从中受益。一个主流的范式是掩码数据建模（masked data modeling），该方法利用可见的数据部分来重建被掩码的部分，从而帮助获得对下游任务有用的表示。然而，传统方法主要关注时间域信息，并对编码器在重建方面提出过高要求，进而削弱了模型的判别能力。本文提出了一种用于心电图分类的双模态掩码自动编码器与内部表示连接（BMIRC）。一方面，BMIRC 在掩码预训练过程中融合了 ECG 的频谱信息，增强了模型对心电图的全面理解；另一方面，它从编码器到解码器建立了内部表示连接（IRC），为解码器提供了多层次的信息以帮助重建，从而让编码器可以专注于建模判别性表示。

关键词：心电图；频谱；双模态；掩码自动编码器；内部表示连接

ABSTRACT

Time series self-supervised methods have been widely used, with electrocardiogram (ECG) classification tasks also reaping their benefits. One mainstream paradigm is masked data modeling, which leverages the visible part of data to reconstruct the masked part, aiding in acquiring useful representations for downstream tasks. However, traditional approach predominantly attends to time domain information and places excessive demands on the encoder for reconstruction, thereby hurting model's discriminative ability. In this paper, we present Bimodal Masked autoencoders with Internal Representation Connections (BMIRC) for ECG classification. On the one hand, BMIRC integrates the frequency spectrum of ECG into the masked pre-training process, enhancing the model's comprehensive understanding of the ECG. On the other hand, it establishes internal representation connections (IRC) from the encoder to the decoder, which offers the decoder various levels of information to aid in reconstruction, thereby allowing the encoder to focus on modeling discriminative representations.

Key Words: ECG; Frequency spectrum; Bimodal; Masked autoencoders; Internal representation connections

目　录

1 引言

1.1 背景介绍

作为一种非侵入性诊断程序，ECG 是检测心律失常最方便和最有效的工具[1]。最初，ECG 分析仅由人类专家执行，这种做法容易受到主观性和不同专业水平的影响，电脑科技的进步，大大提高了心电图的效率和准确性，基于辅助诊断方法，在医学界获得越来越多的关注[2]。

近年来，由于采用了深度学习方法，ECG 分类任务取得了显着进步。研究人员已经通过利用卷积神经网络（CNN），transformer 或其合并展示了令人满意的结果，表明此类方法在该领域的有效性[3-4]。值得注意的是，这些方法主要在监督学习的框架内运行，需要大量的标记数据进行训练。此外，这种训练范式限制了未标记和外生数据源的利用。在与深度学习相关的其他领域，如计算机视觉（CV）和自然语言处理（NLP），研究人员采用自监督学习来解决上述挑战[5-6]。这种范式利用辅助任务从未标记的数据中提取监督信息，使网络能够获得有利于后续任务的表示。一个用自我监督预处理增强的模型，训练从未标记的数据中学习额外的知识，从而与仅通过监督训练训练的模型相比，在监督训练之后实现上级性能。值得注意的是，这些额外的知识可以应用于其他数据集以产生进一步的改进。

最近，用于时间序列分析的自监督方法已经变得流行，ECG 分类任务也从这些进步中获益[7-13]。掩蔽数据建模是该领域的主流范式。与依赖于预定一致性假设的对比学习不同，掩蔽数据建模关注数据的固有特征。这一特征增强了其适用性，从而促使本文将其应用于 ECG 分类任务。然而，这些方法大多只关注时域信息，忽略了来自其他模态或视角的信息，考虑到多模态数据之间的互补性[14]，仅依赖于单一模态的方法无法捕捉到更全面的信息，从而限制了模型的推理和判断能力。在掩蔽数据建模的框架内，当重建目标是原始数据时，最具代表性的范例是掩模自编码器（MAE）。它包括负责编码可见数据的编码器和负责重建掩模数据的解码器。在 ECG 分类任务中，许多心律类别与形态特征相关联。例如，心房颤动（AF）的典型特征是 P 波的缺失。MAE 致力于挖掘原始数据的潜在特征，帮助模型学习这些有区别的细节。这种动力驱使本文利用这种范式来推进本文的研究。

然而，由于自监督重建任务和监督下游任务之间存在差距，MAE 的应用面临一些问题。根据信息瓶颈理论[15]，在典型的监督学习范式下，靠近输入的层捕获更多的低级信息，而靠近输出的层包含更多的高级信息。与上述范式不同，MAE 编码器的浅层和 MAE 解码器的深层都包含丰富的低级信息。此外，各种级别的信

息被保存在网络的中间层中，高级信息通常在编码器的深层中找到[16]。低级信息属于常见的形态细节，它与原始数据密切相关。高级信息属于判别表示，这对于下游分类任务至关重要。解码器将高级表示转换为低级表示依赖于编码器的最终输出。然而，这样的解码过程迫使编码器过度集中于较低级别的信息，旨在帮助解码器更容易地完成重建。简而言之，这种配置导致编码器过度关注重建，从而限制了其学习高级区分表示的能力。

1.2 研究贡献

为了解决上述问题，本文提出了一种新的双峰掩蔽自编码器框架，表示为BMIRC。具体来说，本文采用离散傅立叶变换（DFT）将 ECG 转换为频谱，并将其视为一个独立的模态，这有助于补充模型用于学习的数据源。在频域分析已被广泛应用的先前研究中，这些方法通常涉及直接提取频域特征，随后将其输入到网络中[?]，或者采用双编码器在时域和频域中进行对比表示学习[8]本文利用频谱的方法涉及构建一个双峰联合编码器，旨在学习 ECG 和频谱的联合表示。进行屏蔽数据重构，ECG 和频谱相互补充，促进可转移表示的学习。基于 MAE 范式，本文追求获得更多区分性表示涉及到各级信息的重用。具体来说，从编码器的中间层提取的表示被集成到解码器的各个层中，设计了一种门控表示混合器（GRM）来促进融合。本文将这个过程称为内部表示连接（IRC）。这种方法为解码器提供了各种级别的信息来帮助重建，减轻了解码器的负担，同时鼓励编码器获得更具鉴别力的表示。本文的主要贡献总结如下：

- 本文提出了一种新颖的用于时频联合建模的双模掩蔽自编码器框架。该方法将心电图的频谱集成到掩蔽预训练过程中，使双模联合编码器能够学习全面且通用的表示。
- 本文在编码器和解码器之间建立内部表示连接（IRC），并设计了一个门控表示混合器（GRM）来复用不同层次的信息，从而减轻了解码器的重构负担，同时促进编码器获得更具鉴别力的表示。
- 通过在三个不同的 ECG 数据集上进行的综合实验，BMIRC 在大多数情况下表现出优于竞争基线的上级性能，证明了所提出方法的有效性。

1.3 文章组织结构及章节安排

本文的后续部分结构如下：第 2 节介绍了多模态自监督学习和时间序列自监督学习的相关工作; 第 3 节详细描述了所提出的 BMIRC; 第 4 节概括了本文的结论并描述了未来工作的方向。

2 研究现状

2.1 多模态自监督学习

在单模态研究领域，自监督学习方法取得了显著的进步。因此，当代研究越来越多地将注意力转向多模态领域。当然，值得注意的是，在该领域中，最近的努力的主要部分集中在图像和文本模态的集成上。现有的方法可以大致分为两类。第一类方法集中于通过基于 transformer 的多模态编码器对不同模态之间的交互进行建模。例如，在视觉和语言 Transformer（ViLT）[17]和 Align before transformer（ALBEF）[18]中，来自不同模态的表示用作联合编码器的输入。通过这种融合过程，联合编码器在不同的预训练任务中进行训练，第二类方法强调为单个模态训练专用编码器。例如，对比度图像预训练（CLIP）[19]利用对比度损失对编码器施加约束，促进跨不同模态的一致表示的获取。

虽然这些方法改进了图像和文本模态的表示学习，但对于具有独特属性和有限数据的时间序列，它们可能不太有效。在心电图分析的背景下，从其他模态同时收集的数据的稀缺性对多模态自监督方法的适用性构成了限制。此外，为图像和文本模态设计的预训练任务的有效性在应用于心电图和其他相关模态时仍然不确定。因此，本文致力于从原始 ECG 的不同角度探索数据源，并设计基于屏蔽数据建模的双峰自监督方法，这提供了更强的适用性。

2.2 时间序列的自监督学习

时间序列自监督学习方法的日益流行主要是由于 CV 和 NLP 的实质性进展。

一种范式是基于对比学习，其重点是通过施加不变性约束来进行表征学习。例如，时频一致性（TF-C）[8]假定时频域内的一致性，并通过减少相同样本的时频域表示之间的差异来进行预训练。对比预测编码（CPC）[11]和时频一致性（T-C）[13]通过时间和上下文对比的系列表示学习框架（TS-TCC）[7]主要强调时间不变性。CPC 通过预测后续时间步的表示来实现它，而 TS-TCC 则致力于最大化跨不同视图的相同时间跨度的表示之间的相似性。通用时间序列表示学习（TimesURL）[20]首先引入了一种基于频率-时间的增强来保持时间属性。值得注意的是，时间重建作为一个联合优化目标与对比学习相结合，以捕获片段级别和实例级别的信息。然而，必须承认，无论是不同的增强策略还是不变性假设都不适用于所有场景。

另一种范式依赖于屏蔽数据建模，利用自编码器在可见数据的基础上重构屏蔽数据，其代表作是 Patch Time Series Transformer（PatchTST）[9]和时间序列掩蔽

自动编码器（TimeMAE）[10]，共享一个共同的特性，因为它们都将时间序列划分为非重叠的补丁，PatchTST 遵循 MAE 范式，旨在重建原始时间序列。相比之下，TimeMAE 致力于特征级重建。此外，有一种方法认为，随机掩蔽某些时间点的标准方法可能会严重破坏时间序列中重要的时间变化。用于掩蔽时间序列建模的简单预训练框架（SimMTM）[21]通过加权聚合流形外的多个邻居来恢复被屏蔽的时间点。这些方法不依赖于人为预定的假设，因此更适合在新领域应用。

受此鼓舞，本文采用屏蔽数据建模的范式来制定本文的模型。在 ECG 表示学习领域，CPC 在获取广义表示方面表现出有效性[12]。基于 MAE 范式的几种变体被提出来在时间和变量的双重维度中执行掩蔽数据建模[13]。然而，这些方法目前仅限于直接应用和对其他领域现有方法的简单改进，在屏蔽数据建模的基础上，引入心电图的频谱特征进行双峰学习，并通过内部表示连接（IRC）鼓励编码器学习更多有区别的表示，以获得更全面、更有区别的表示，从而提高下游分类任务的性能。

3 方法

以下部分详细描述了本文提出的方法，如图3.1所示。

本文的方法包括预训练和微调阶段。由于数据集只提供原始 ECG，首先生成 ECG 和频谱的补丁嵌入，表示为时频补丁嵌入。在预训练阶段，时频补丁嵌入在掩蔽之后被馈送到双峰联合编码器。然后，该算法采用模态相关解码器，在 IRC 的帮助下重建被屏蔽的数据，该阶段不需要任何标签的参与，在微调阶段，解码器由分类器代替，分类器由全局平均池化层和线性层组成.

剩余的网络使用预训练的模型参数进行初始化。这个阶段需要进行常规的监督训练。更多细节如图 2 所示。接下来，本文将介绍上述步骤的每个组成部分。

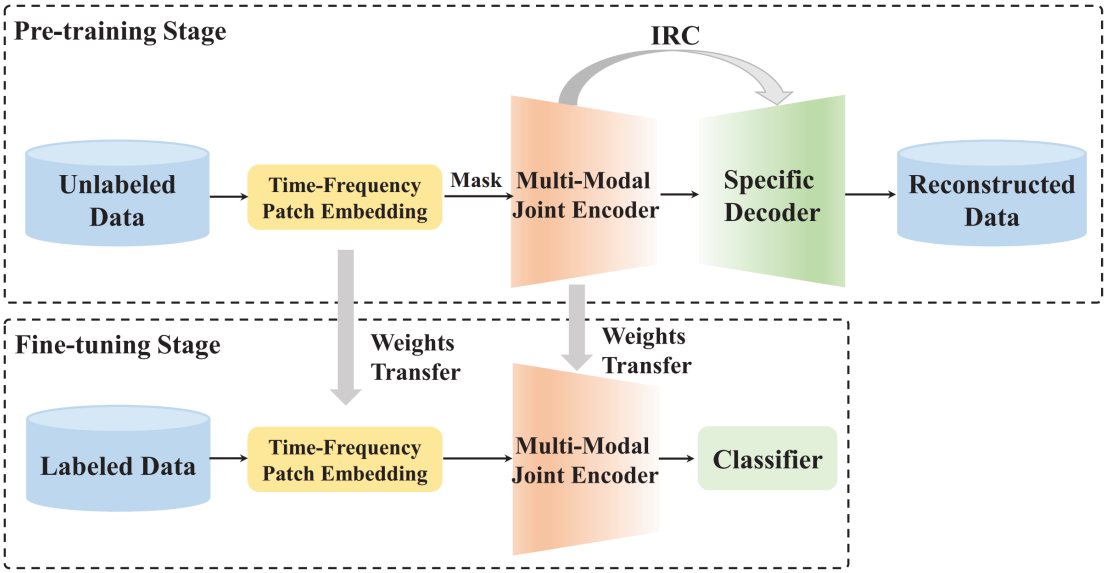

图 3.1　BMIRC 的整个过程。它由两个阶段组成：预培训和微调。

3.1 时频贴片嵌入

在获得最终解码器层 U_H^m 的输出后，通过层归一化-线性模块 P_m 和重塑操作将其转换为与时间或频率模态的维度匹配（$\mathbb{R}^{C\times L}$ 或 $\mathbb{R}^{C\times \frac{L}{2}}$），即：

正如图3.2所述，一个心电图（ECG）的第 i 导联，表示为 $t_i \in \mathbb{R}^L$，通过离散傅里叶变换（DFT）被转换为频谱 $f_j \in \mathbb{R}^N$：

$$f_j(k) = \mathrm{DFT}[t_i] = \sum_{n=0}^{N-1} t_i(n) \cdot e^{-j\frac{2\pi}{N}kn}, k = 0, 1, \ldots, N-1$$

其中，n 是时间点的索引，k 表示 ECG 中存在的不同频率。根据频域采样定理，DFT 变换的区间 N 必须大于或等于 ECG 的长度 L，以防止混叠。在这种情况下，

本文设定 $N = L$ 来得到 $f_j \in \mathbb{R}^L$。这一过程可以通过快速傅里叶变换（FFT）算法有效计算。

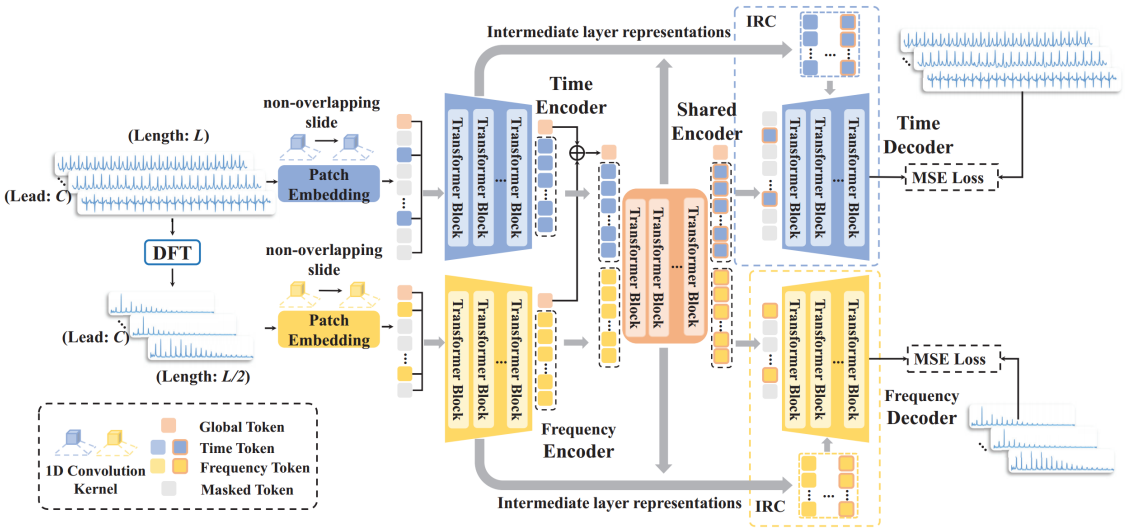

图 3.2　预训练阶段的整体框架

由于在变换后的频谱中存在对称性，本文截取了 f_j 的前半部分 $f_j^* \in \mathbb{R}^{\frac{L}{2}}$，以减少冗余。频谱揭示了数据中存在的各种频率成分的比例，从而从另一个角度揭示数据的特性。各种类型的 ECG 在时间域中展现出不同的形态特征。例如，房颤（AF）的典型特征是 P 波的缺失，而左束支阻滞（LBBB）和右束支阻滞（RBBB）则通过 QRS 波形的改变来区分。ECG 中存在的变化也反映在频谱中。这一事实直观地支持了整合频域信息来增强模型判别能力的合理性。

在大多数方法以逐点方式处理时间序列数据的背景下，基于 patch 的建模方法已被证明是有效的。特别是，推断被掩码的区域比推断被掩码的点更具挑战性。因此，基于 patch 的重建任务鼓励预训练模型学习更多潜在信息。在本文的方法中，ECG 和频谱被分成不重叠的 patch 以进行编码。一个多导联 ECG 及其对应的频谱表示为 $T = [t_1, t_2, \ldots, t_C] \in \mathbb{R}^{L \times C}$ 和 $F = [f_1^*, f_2^*, \ldots, f_C^*] \in \mathbb{R}^{\frac{L}{2} \times C}$，其中 C 是导联的数量。本文使用两个一维卷积层对来自两种模态的 patch 进行编码。卷积核的大小设为 $S \times C$，步长设为 S，以确保 patch 的独立性。在这种配置下，每个 patch 的长度为 S，表示心电图的一个片段。按照 MAE 的思想，patch 嵌入被表示为 token，每个 token 对应于特定 patch 的嵌入。T 和 F 的 token 表示为：

$$Z_t = [z_t^1, z_t^2, \ldots, z_t^{\frac{L}{S}}] \in \mathbb{R}^{\frac{L}{S} \times D}, Z_f = [z_f^1, z_f^2, \ldots, z_f^{\frac{L}{2S}}] \in \mathbb{R}^{\frac{L}{2S} \times D}$$

其中，D 表示卷积核的数量，表示每个 token 的维度。”t” 和”f” 分别表示 ECG 和

频谱模态。为了便于表述，ECG 和频谱也分别被称为时间和频率模态。在接下来的部分中，使用了 $N_t = \frac{L}{S}$ 和 $N_f = \frac{L}{2S}$ 来表示时间和频率模态中的 token 数量。

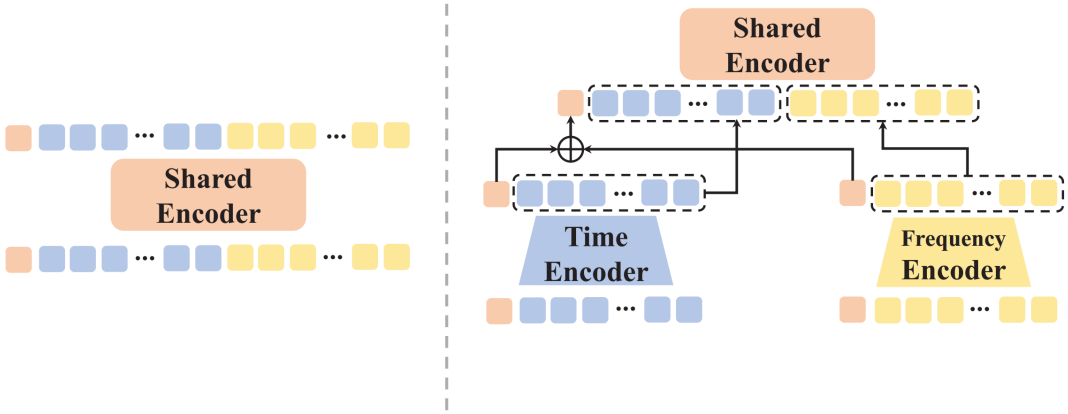

图 3.3　不同的双模 Transformer 架构

3.2　双模联合编码器

在本文的方法中，Transformer 作为编码器的主要组件。鉴于自注意力机制对输入位置的天然不敏感性，可学习的位置嵌入 PE $\in \mathbb{R}^{N \times D}$ 被集成到 patch 嵌入中，以增强模型的 token 定位能力。此外，还为每个模态引入了一个额外的可学习全局 token $z_g \in \mathbb{R}^D$，其中"g" 表示" 全局"，从而促进全局信息的提取。最后，对于模态 $m \in \{t, f\}$，输入 tokens $I_m \in \mathbb{R}^{N_m \times D}$ 表示为：

$$\tilde{I}_m = Z_m + \mathrm{PE}_m$$

$$I_m = \mathrm{Concat}(z_g^m, \tilde{I}_m)$$

其中，Concat 是连接运算符，“t” 和 “f” 分别代表时间模态和频率模态。

3.2.1　遮掩策略 (Masking Strategy)

在预训练阶段，输入到编码器的 tokens 需要以预定速率进行遮掩。受 MAE 启发，本文采用了一种随机遮掩策略，意味着每个 token 有相同的概率被遮掩。由于每次遮掩操作是随机的，在各个批次和训练轮次之间，重构任务的表现具有多样性，从而使预训练任务更加具有挑战性。高遮掩率，例如 60% 和 75%，表示重构

难度较高，并已被证明可以提升预训练模型的性能[6,10]。然而，由于时间模态和频率模态的数据特性不同，在联合建模中采用统一的遮掩率并不理想。通过实验，本文确定遮掩率为 50%（用于 ECG）和 75%（用于频率谱）能帮助模型实现最佳性能。考虑到 ECG 相比频率谱的复杂性更高，采用相对较低的遮掩率有助于提升模型性能，这凸显了在建模能力和重构难度之间权衡的必要性。

3.2.2 总体结构

本文的双模态联合编码器由两部分组成：模态特定编码器（时间编码器和频率编码器）和共享编码器。前者专注于在每个模态中建模表示，而后者负责捕捉两个模态之间的交互。两部分都使用 Transformer 模块进行构建。本文将时间和频率模态的输入 tokens 馈入模态特定编码器 E_m，以获得输出表示 O_m。

$$O_m = E_m(I_m)$$

其中 $m \in \{t, f\}$ 表示时间或频率模态。

在输入共享编码器之前，各个模态的 tokens 经过初步融合，经过层归一化（LN）处理。具体来说，时间和频率模态的全局 tokens 相加，并插入到序列的第一个位置，其他 tokens 依次串联。然后，使用共享编码器促进双模态表示的深度融合。形式上，模态特定编码器最终层的输出表示表示为 $O_m = [o_g^m, o_1^m, o_2^m, \ldots, o_n^m]$，本文将其传递到共享编码器 Θ 中，按照以下方程式处理：

$$\tilde{O}_m = LN(O_m) = [\tilde{o}_g^m, \tilde{o}_1^m, \tilde{o}_2^m, \ldots, \tilde{o}_n^m]$$

$$O_0^s = [o_g^t + \tilde{o}_g^f, \tilde{o}_1^t, \tilde{o}_2^t, \ldots, \tilde{o}_n^t, \tilde{o}_1^f, \tilde{o}_2^f, \ldots, \tilde{o}_n^f]$$

$$O_s = \Theta(O_0)$$

其中 O_s 是共享编码器的输出表示。

本文对模型架构进行了消融实验，比较了本文的架构（如图3.3(b) 所示）与仅使用共享编码器建模双模态表示的替代架构（如图3.3(a) 所示）。

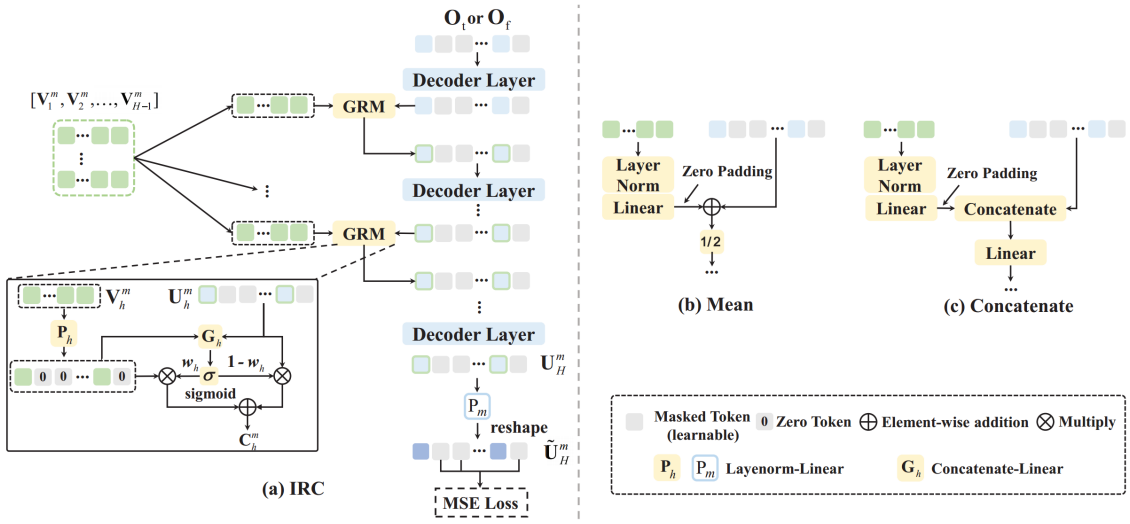

图 3.4　内部表示连接（IRC）的结构及其各种融合方法

3.3　内部表示连接（Internal Representation Connections）

鉴于时间模态和频率模态之间的明显差异，特定于模态的解码器（时间解码器和频率解码器）被设计用于执行重构任务。它们都是由 transformer 块组成的。为了减少计算成本，使用了浅层解码器，它们的维度和层数比编码器少。He 等人[6]已经证明这种设计不会影响模型的性能，而且更深的解码器并不意味着更好的性能。这与仅在 MAE 范式中，只有来自编码器最终层的表示被输入到解码器中不同，本文为解码器提供了来自编码器中间层的更多表示，以帮助解码器逐步完成重构任务。因此，内部表示连接（IRC）从编码器到解码器被建立起来。

IRC 缓解了解码器的重构压力，从而防止了编码器因优先重构而承受不必要的负担。这样，编码器就能够更好地获得高级别的区分性表示。

共享编码器的输出表示 O_s 被分割为时间模态的 O_t 和频率模态的 O_f。O_t 和 O_f 首先通过层规范化-线性模块转换以减少维度。随后，它们与可学习的掩码标记（masked tokens）连接，并输入到具有可学习位置嵌入的模态特定解码器中。值得注意的是，IRC 不存在于解码器的第一层中。

3.3.1　门控表示混合器（GRM）

在解码器的深度为 H 的情况下，本文从编码器中均匀地选择 $H-1$ 层表示 $[V_h^m, V_{h+1}^m, \ldots, V_H^m]$ 进行融合，其中 m 代表模态（时间模态或频率模态），小下标代表来自较深层次的表示。为了融合这些编码器的表示，本文设计了一个叫做门控表示混合器（GRM）的机制。

正如图3.4（a）所示，针对模态 m，在解码器的第 h 层，来自编码器的第 h 层表示 V_h 通过门控机制与解码器的第 h 层表示 U_h^m 融合。层规范化-线性模块（由 P_h 表示）用于将 V_h 转换为 $\hat{V}_h$，从而促进需要融合的表示之间的对齐。由于解码器输出 U_h^m 包含可学习的掩码标记，本文使用零标记填充 $\hat{V}_h$ 的对应位置，以保持维度的一致性。

门控单元 G_h 包含连接和线性变换，用来控制输入表示到输出的贡献。由于只有两个输入，Sigmoid 激活函数 σ 被用来将贡献转换为相应的权重向量 $w_h \in \mathbb{R}^N$，这意味着每个 token 都有一个融合权重。在 token 层级，$\hat{V}_h$ 和 U_h^m 经过加权后相加，权重由 w_h 引导，这意味着可以对每个 token 应用自适应融合策略。随后，融合后的表示 C_h^m 被输入到第 $h+1$ 层的 Transformer 块 Λ_{h+1} 以获得 U_{h+1}^m。

特别是，以上描述的融合仅涉及可见 token，旨在补充各种信息的不同层次，促进掩码 token 学习过程的进展。该过程在每个后续层中迭代进行，编码器的浅层表示逐步融合到解码器的更深层，进而逐渐推动重构过程的进展。上述过程通过以下公式表示：

$$\hat{V}_h = P_h(V_h)$$

$$w_h = \sigma(G_h(\hat{V}_h, U_h^m))$$

$$C_h^m = w_h * \hat{V}_h + (1 - w_h) * U_h^m$$

$$U_{h+1}^m = \Lambda_{h+1}(C_h^m)$$

如图3.4(b) 和 (c) 所示，其他两种替代方法如下：1）均值法：通过层归一化-线性模块对齐后，输出平均值；2）拼接法：通过层归一化-线性模块对齐后，首先对输入表示进行拼接，然后使用线性层统一维度。在上述两种融合方法中，应用零填充来对齐掩码标记。

3.3.2 双模态重构损失

在获得最终解码器层 U_H^m 的输出后，通过层归一化-线性模块 P_m 和重塑操作将其转换为与时间或频率模态的维度匹配（$\mathbb{R}^{C\times L}$ 或 $\mathbb{R}^{C\times \frac{L}{2}}$），即：

$$\tilde{U}_H^m = Reshape(P_m(U_H^m)),\ m \in \{t, f\}$$

遵循 MAE（掩码自动编码器），重构损失仅在掩码标记上进行计算：

$$\mathcal{L}_{time} = MSE(U_H^t, T_{masked})$$

$$\mathcal{L}_{freq} = MSE(U_H^f, F_{masked})$$

$$\mathcal{L}_{recon} = \alpha\mathcal{L}_{time} + \beta\mathcal{L}_{freq}$$

其中 MSE 为均方误差，T_{masked} 和 F_{masked} 分别代表时间和频率模态中掩码标记的真实值，α 和 β 代表两个模态重构损失的权重。

4 结论

本文提出了一种新的用于心电图分类的双峰掩蔽自编码器框架 BMIRC。BMIRC 的一个特点是将心电图产生的频谱作为一个独立的模态纳入掩蔽预训练过程。所提出的双峰联合编码器捕获从模态内到模态间的相互作用，从而获得更全面的表示。此外，本文建立了从编码器到解码器的内部表示连接（IRC），并设计了一个门控表示混合器（GRM）来执行融合过程。这些增强减轻了解码器的负担，从而增强编码器学习判别表示的能力。本文在域内和跨域内进行了多组实验，在三个公开的 ECG 数据集上进行了实验，结果证明了 BMIRC 的有效性，并进一步证明了预训练的 ECG 权值的可移植性

然而，本文模型内部特征的可解释性仍然是一个挑战，这也是基于深度学习方法的痛点。在医疗领域的实际应用中，模型决策过程的可解释性对于临床医生来说尤为重要。未来，本文将致力于开发新的方法来提高 BMIRC 的可解释性，在模型的内部细节方面，本文计划进一步探索和优化跨模态信息融合技术，以减少不同模态之间的不一致性和噪声，从而提高模型的稳定性和鲁棒性。

参考文献

[1] HONG S, ZHOU Y, SHANG J, et al. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review [J]. Computers in Biology and Medicine, 2020, 122: 103801.

[2] SAHOO S, DASH M, BEHERA S, et al. Machine learning approach to detect cardiac arrhythmias in ecg signals: A survey [J]. IRBM, 2020, 41(4): 185-194.

[3] LIU X, WANG H, LI Z, et al. Deep learning in ecg diagnosis: A review [J]. Knowledge-Based Systems, 2021, 227: 107187.

[4] XIA Y, XIONG Y, WANG K. A transformer model blended with cnn and denoising autoencoder for inter-patient ecg arrhythmia classification [J]. Biomedical Signal Processing and Control, 2023, 86: 105271.

[5] Chen T, Kornblith S, Norouzi M, et al. A Simple Framework for Contrastive Learning of Visual Representations [A]. 2020: arXiv:2002.05709. arXiv: 2002.05709.

[6] He K, Chen X, Xie S, et al. Masked Autoencoders Are Scalable Vision Learners [A]. 2021: arXiv:2111.06377. arXiv: 2111.06377.

[7] Eldele E, Ragab M, Chen Z, et al. Time-Series Representation Learning via Temporal and Contextual Contrasting [A]. 2021: arXiv:2106.14112. arXiv: 2106.14112.

[8] Zhang X, Zhao Z, Tsiligkaridis T, et al. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency [A]. 2022: arXiv:2206.08496. arXiv: 2206.08496.

[9] Nie Y, Nguyen N H, Sinthong P, et al. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers [A]. 2022: arXiv:2211.14730. arXiv: 2211.14730.

[10] Cheng M, Liu Q, Liu Z, et al. TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders [A]. 2023: arXiv:2303.00320. arXiv: 2303.00320.

[11] van den Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding [A]. 2018: arXiv:1807.03748. arXiv: 1807.03748.

[12] Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data [A]. 2021: arXiv:2103.12676. arXiv: 2103.12676.

[13] ZHANG H, LIU W, SHI J, et al. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-15.

[14] Baltrušaitis T, Ahuja C, Morency L P. Multimodal Machine Learning: A Survey and Taxonomy [A]. 2017: arXiv:1705.09406. arXiv: 1705.09406.

[15] Tishby N, Zaslavsky N. Deep Learning and the Information Bottleneck Principle [A]. 2015: arXiv:1503.02406. arXiv: 1503.02406.

[16] Han Q, Cai Y, Zhang X. RevColV2: Exploring Disentangled Representations in Masked Image Modeling [A]. 2023: arXiv:2309.01005. arXiv: 2309.01005.

[17] Kim W, Son B, Kim I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision [A]. 2021: arXiv:2102.03334. arXiv: 2102.03334.

[18] Li J, Selvaraju R R, Gotmare A D, et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation [A]. 2021: arXiv:2107.07651. arXiv: 2107.07651.

[19] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision [A]. 2021: arXiv:2103.00020. arXiv: 2103.00020.

[20] Liu J, Chen S. TimesURL: Self-supervised Contrastive Learning for Universal Time Series Representation Learning [A]. 2023: arXiv:2312.15709. arXiv: 2312.15709.

[21] Dong J, Wu H, Zhang H, et al. SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling [A]. 2023: arXiv:2302.00861. arXiv: 2302.00861.

[22] KNUTH D E. Literate programming [J]. The Computer Journal, 1984, 27(2): 97-111.

[23] LESK M, KERNIGHAN B. Computer typesetting of technical journals on UNIX [C]. Proceedings of American Federation of Information Processing Societies: 1977 National Computer Conference. Dallas, Texas, 1977: 879-888.

[24] LE M D, SINGH RATHOUR V, TRUONG Q S, et al. Multi-module recurrent convolutional neural network with transformer encoder for ecg arrhythmia classification [C]. 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). 2021: 1-5.